Overcoming Barriers to Skill Injection in Language Modeling: Case Study in Arithmetic

Mandar Sharma
Virginia Tech
mandarsharma@vt.edu

Nikhil Muralidhar
Stevens Institute of Technology
nmurali1@stevens.edu

Naren Ramakrishnan
Virginia Tech
naren@cs.vt.edu

Abstract

Through their transfer learning abilities, highly-parameterized large pre-trained language models have dominated the NLP landscape for a multitude of downstream language tasks. Though linguistically proficient, the inability of these models to incorporate the learning of non-linguistic entities (numerals and arithmetic reasoning) limits their usage for tasks that require numeric comprehension or strict mathematical reasoning. However, as we illustrate in this paper, building a general purpose language model that also happens to be proficient in mathematical reasoning is not as straight-forward as training it on a numeric dataset. In this work, we develop a novel framework that enables language models to be mathematically proficient while retaining their linguistic prowess. Specifically, we offer information-theoretic interventions to overcome the catastrophic forgetting of linguistic skills that occurs while injecting non-linguistic skills into language models.

1 Introduction

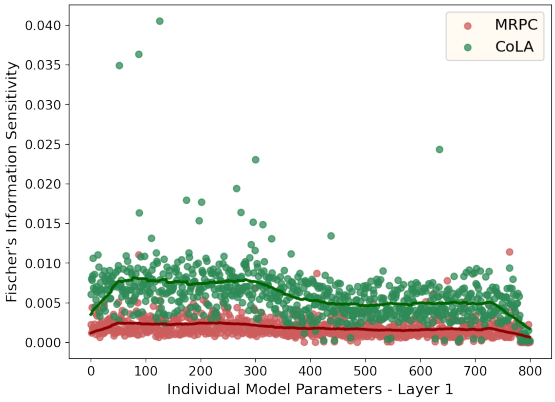
Numerals grant objectivity to language [15], thus, their incorporation to language models, among other abilities, allows for better information extraction and language inference [12, 14, 18]. With the rise of the Math-NLP niche, several notable publications have investigated the deficiency of inherent numerical skills induced in large language models through unsupervised training [23, 8, 27]. Several more offer architectures and training schemes to induce this skill through supervision [19, 7, 5, 20]. However, the overarching goal should remain to build numerically-capable language models that still do what they were intended to do - *model language*.

Injecting non-linguistic skills often comes at the cost of losing linguistic skills. Akin to all neural models, language models are susceptible to catastrophic forgetting [10] when trained for multi-task learning, especially when the two tasks are substantially different - linguistic vs non-linguistic. Here, we first investigate the nature of catastrophic forgetting in the context of language models through an information-theoretic lens and subsequently establish interventions that prevent this phenomenon. Our models perform substantially better arithmetic while retaining their linguistic prowess.

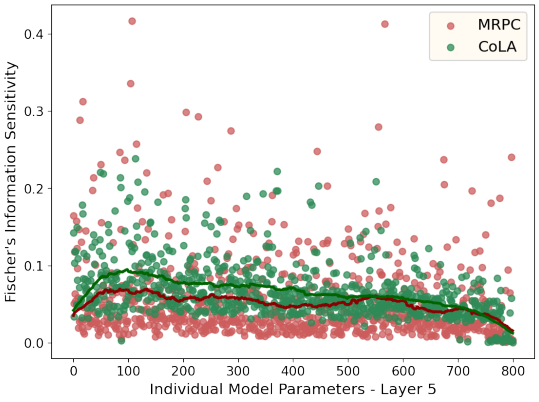
2 Catastrophic Forgetting in Language Models

In multi-task learning, when a model that is pre-trained on task A is subsequently trained on task B, the weights in the model that are vital for task A adapt their values to meet the objectives of task B. It has been found that this frequently leads to performance degradation on task A. This phenomenon is referred to as catastrophic forgetting [10]. While it has been found that off-the-self pre-trained language models do establish a notion of numeric scale from their unsupervised pre-training, these are not quite enough to perform commonsense reasoning [27]. Furthermore, without additional interventions, these notions of numeric scales fail to extrapolate for numerals outside of the training set [23]. To meet the numerical comprehension standards required for commonsense reasoning, these model have to be trained further on numeric datasets, often resulting in the loss of their original linguistic prowess through catastrophic forgetting.

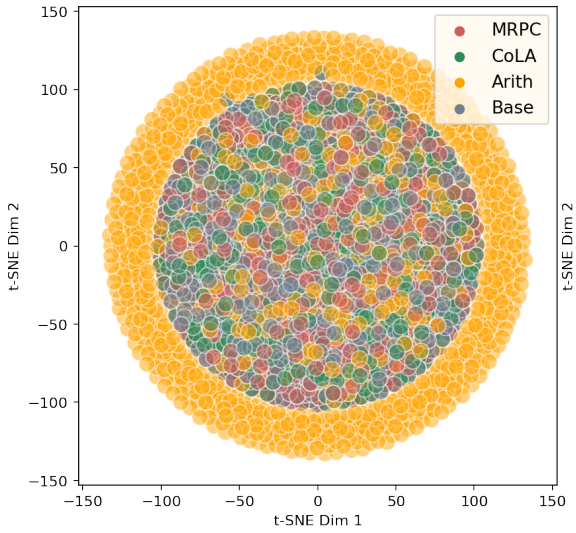
36th Conference on Neural Information Processing Systems (NeurIPS 2022) Workshop on Math-AI.



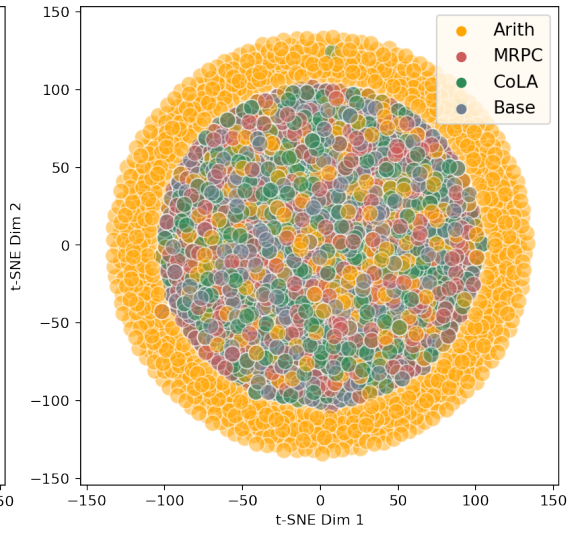
(a) Information Sensitivity
Self-attention Layer-1



(b) Information Sensitivity
Self-attention Layer-5



(c) Encoder Parameter Space
Self-attention Layer-1



(d) Encoder Parameter Space
Self-attention Layer-5

Figure 1: Parameter sensitivity (a & b) for the $n = 800$ most vital parameters of the encoder layers 1 & 5 for CoLA and MRPC. The higher sensitivity of these parameters to CoLA vs MRPC lends an information-theoretic explanation of task-specific performance degradation. 2D t-SNE visualizations of the parameterization space (c & d) for the same models & layers - the distinction between the parameterization space of models trained for Arithmetic task vs Linguistic tasks offers another perspective into degradation of lingusitic performance when trained on non-linguistic task.

Interestingly, for langauge models, the catastrophic forgetting tendencies are not evenly spread across standardized GLUE [24] tasks - for instance, a base DistilBERT model [17] trained for arithmetic computation faces significant degradation in its grammatical prowess (CoLA [25]) while still retaining much of its ability for semantic equivalence (MRPC [3]), as seen in the second row of table 1.

In an effort to better understand this phenomenon, we adopt a two-pronged approach. The analysis of the weights and biases of a converged model across multiple datasets provides a parameterization-standpoint perspective towards understanding their performance across shared tasks [6, 21]. Thus, we visualize the parameterization space for models trained independently on each of the aforementioned three tasks through figures 1c and 1d. Through the low-dimensional t-SNE [22] projections, we observe that the model parameterizations live in different spaces when the tasks are linguistic (CoLA and MRPC) vs non-linguistic (Arithmetic). This ties back to the premise of catastrophic forgetting - where the parameters of the DistilBERT model trained for arithmetic computation have re-adjusted their weights to a space that does not comply well with linguistic tasks. These new parameter

distributions for arithmetic computation (orange) are consistent across the transformer encoder self-attention layers - as seen for layer 1 in figure 1c and layer 5 in figure 1d. These two figures offer the first perspective on linguistic vs non-linguistic performance degradation from a model parameterization stand-point.

For a complementary perspective, we zoom into the contributions of the individual model parameters for our set of shared tasks through an information-theoretic lens. For a single sample Y drawn from a distribution with probability desnity $f(y;\theta)$, the Fisher information index $I(\theta)$ (1) quantifies the sensitivity of the parameter θ to the data instance Y. After selecting the top $n = 800$ most vital parameters for each encoder layer for the Arithmetic task based on their Fisher information scores, we observe that the same model parameters are more sensitive to the CoLA task than to the MRPC task, as shown in figures 1a and 1b, again, based on their respective task-specific sensitivity scores. Thus, offering a complementary information-theoretic perspective on task-specific performance degradation.

$$I(\theta) = E(\frac{d \log f(Y;\theta)}{d\theta})^2 = -E(\frac{d^2 \log f(Y;\theta)}{d\theta^2}) \tag{1}$$

3 Overcoming Catastrophic Forgetting

3.1 Elastic Weight Consolidation for Language Modeling

System-level consolidation often consists of stitching-together an amalgamated dataset that consistutes of multiple-shared tasks [11]. However, for general-purpose language models, the range of possible downstream tasks are so diverse that the paradigm has remained to build large models that hold linguistic prowess and are intended to be fine-tuned on a single downstream task [17, 9, 1]. Thus, being more suited to a continual learning paradigm. As these models, by their stature, are highly parameterized - it can be ascertained that there is a solution for task B, θ_B, that is proximal to the solution space for task A, θ_A. For such continual learning, Elastic Weight Consolidation (EWC) [10] regularizes learning on specific network weights based on their importance to previously seen tasks, ensuring θ_B remains proximal to θ_A.

Thus, with the posterior distribution of the DistilBERT model approximated through its Fischer information matrix F and Gaussian distribution mean from parameters θ^*_{gen}, we train the model to predict the correct tokens representing the results of arithmetic operations through masked-modeling loss $\mathcal{L}_{arith}$ such that changes to any model parameter i crucial to then general functionality of the model θ_{gen} is penalized through a quadratic penalty scaled by λ (2). The selection of the hyperparameter λ is crucial as it dictates both model convergence and balances the learning of θ_{arith} with θ_{gen}. Thus, we gauge the sensitivity of model convergence with respect to λ with a hyperparameter sweep raning from λ=1e-4 to 5e-11. Figure 2 shows the interplay between the weight consolidation loss EWC and the Cross-Entropy loss for learning arithmetic CE (color-matched) for different values of λ. We observe the first sign of model convergence at λ=1e-8, and although the model convergence improves with decreasing values of λ, we set λ to 1e-8 for our experiments to promote balanced learning of θ_{arith} with proper retention of θ_{gen}.

$$\mathcal{L}(\theta) = \mathcal{L}_{arith}(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{gen,i})^2 \tag{2}$$

3.2 Experimental Setup and Results

Our in-house arithmetic dataset comprises of numerals modeled after the numeral distributions in real-world datasets DROP [4] and EQAUTE [16] (§Appendix 5.2) with 21,838 instances of arithmetic computations. All models are trained for 50 epochs where the datasets are processed through the standard sub-word tokenization scheme of BERT [9]. These models are then fine-tuned and evaluated on the range of GLUE tasks [24]:

- Single Sentence Tasks: Corpus of Linguistic Acceptability (CoLA) [25] for grammatical fidelity with Matthews correlation coefficient as the metric and the Stanford Sentiment Treebank (SST-2) [13] for sentiment analysis with accuracy as the metric.

- Similarity and Paraphrasing Tasks: The Microsoft Research Paraphrase Corpus (MRPC) [3] for semantic equivalence with averaged accuracy and f1 scores as the metric and the Semantic Textual Similarity Benchmark (STS-B) [2] for sentence similarity with the Spearman rank correlation coefficient as the metric.
- Inference Task: The Multi-genre Natural Language Inference Corpus (MNLI) [26] for textual entailment between a given premise and hypothesis with accuracy as the metric.

The results presented in table 1 represent task-metrics for the models as μ_σ where μ represents the mean value and σ represents the standard deviation across two runs with different seed values for random initialization of the model weights. Please note that the base DistilBERT model has been used off-the-shelf and thus has no standard deviation across its runs.

Table 1: Comparative analysis between the base model, the base model trained on arithmetic, and our model. The results of the arithmetic computations are measured based on the *ln* RMSE of the model output and the ground-truth numeral. The results of the GLUE tasks follow the metrics as described in §3.1.

	ln RMSE	CoLA	MNLI	MRPC	SST-2	STS-B
Base	$3.5367_{0.000}$	$\mathbf{0.4827}_{0.000}$	$\mathbf{0.8074}_{0.000}$	$\mathbf{0.8797}_{0.000}$	$\mathbf{0.8967}_{0.000}$	$\mathbf{0.8740}_{0.000}$
Base + Arithmetics	$0.4443_{0.011}$	$0.0000_{0.000}$	$0.3553_{0.001}$	$0.7524_{0.000}$	$0.8761_{0.003}$	$0.3998_{0.079}$
Ours	$\mathbf{0.4360}_{0.016}$	$0.4193_{0.000}$	$0.7951_{0.004}$	$0.8570_{0.000}$	$0.8962_{0.008}$	$0.8626_{0.005}$

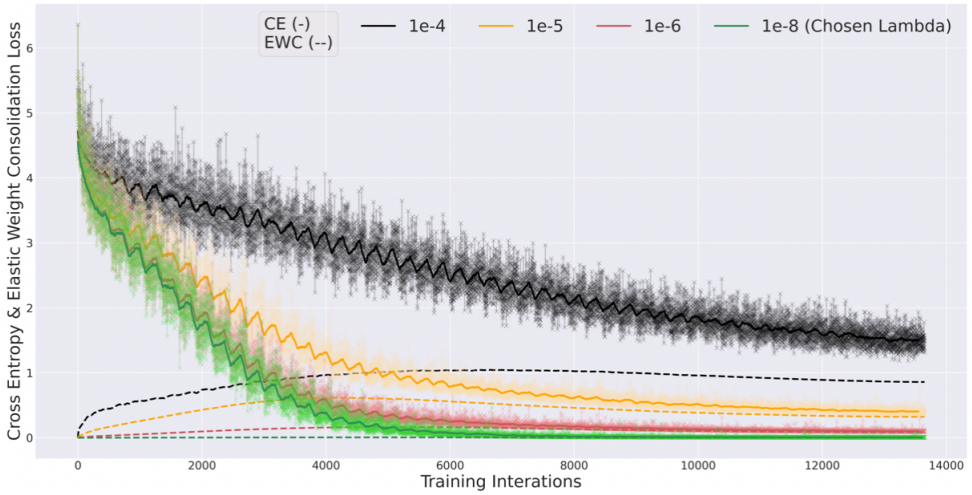


Figure 2: Interplay between the CE & the EWC loss (color-matched) as a function of training iterations. The first sign of model convergence is observed at λ=1e-8, and although the convergence improves with decreasing values of λ, we choose λ to 1e-8 to promote balanced learning of θ_{arith} with proper retention of θ_{gen}.

4 Discussion

From the results in table 1, we infer that our model, while performing the best in arithmetic computation - orders of magnitude better than the base model, closely competes with the linguistic prowess of the base model. Besides preventing performance degradation on the MNLI, STS-B, and the MNLI tasks, for the CoLA task that suffered significant degradation with its Matthews correlation coefficient dropping to 0 after being trained on the arithmetic dataset, our model was able to revitalize it to a value competitive to the base model.

Returning to our original premise, this paper serves as a proof-of-concept that the inherent barries to skill injection caused by the catastrophic forgetting tendencies of large networks can be overcome with weight consolidation schemes specifically tailored for language modeling. Please see Appendix §5.1 for a disucssion on the limitations of our efforts. Our datasets as well as the codebase is hosted at `https://github.com/Mandar-Sharma/OvercomingBarriers`.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[2] Daniel Cera, Mona Diabb, Eneko Agirrec, Inigo Lopez-Gazpioc, Lucia Speciad, and Basque Country Donostia. Semeval-2017 task 1: Semantic textual similarity multilingual and cross-lingual focused evaluation.

[3] Bill Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*, 2005.

[4] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, 2019.

[5] Mor Geva, Ankit Gupta, and Jonathan Berant. Injecting numerical reasoning skills into language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, 2020.

[6] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in Neural Information Processing Systems*, 28, 2015.

[7] Chengyue Jiang, Zhonglin Nian, Kaihao Guo, Shanbo Chu, Yinggong Zhao, Libin Shen, and Kewei Tu. Learning numeral embedding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2586–2599, 2020.

[8] Devin Johnson, Denise Mak, Andrew Barker, and Lexi Loessberg-Zahl. Probing for multilingual numerical understanding in transformer-based language models. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 184–192, 2020.

[9] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[10] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[11] Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20(7):512–534, 2016.

[12] Aman Madaan, Ashish Mittal, Ganesh Ramakrishnan, Sunita Sarawagi, et al. Numerical relation extraction with minimal supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[13] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. *Advances in neural information processing systems*, 30, 2017.

[14] Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, 2018.

[15] Theodore M Porter. Trust in numbers. In *Trust in Numbers*. Princeton University Press, 1996.

[16] Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. Equate: A benchmark evaluation framework for quantitative reasoning in natural language inference. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 349–361, 2019.

[17] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled bersion of bert: Smaller, faster, cheaper and lighter. 2019.

[18] Mandar Sharma, Ajay Gogineni, and Naren Ramakrishnan. Innovations in neural data-to-text generation. *arXiv preprint arXiv:2207.12571*, 2022.

[19] Georgios Spithourakis and Sebastian Riedel. Numeracy for language models: Evaluating and improving their ability to predict numbers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2104–2115, 2018.

[20] Dhanasekar Sundararaman, Shijing Si, Vivek Subramanian, Guoyin Wang, Devamanyu Hazarika, and Lawrence Carin. Methods for numeracy-preserving word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4742–4753, 2020.

[21] Thomas Unterthiner, Daniel Keysers, Sylvain Gelly, Olivier Bousquet, and Ilya Tolstikhin. Predicting neural network accuracy from weights. 2020.

[22] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

[23] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do nlp models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5307–5315, 2019.

[24] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018.

[25] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Cola: The corpus of linguistic acceptability. 2019.

[26] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, 2018.

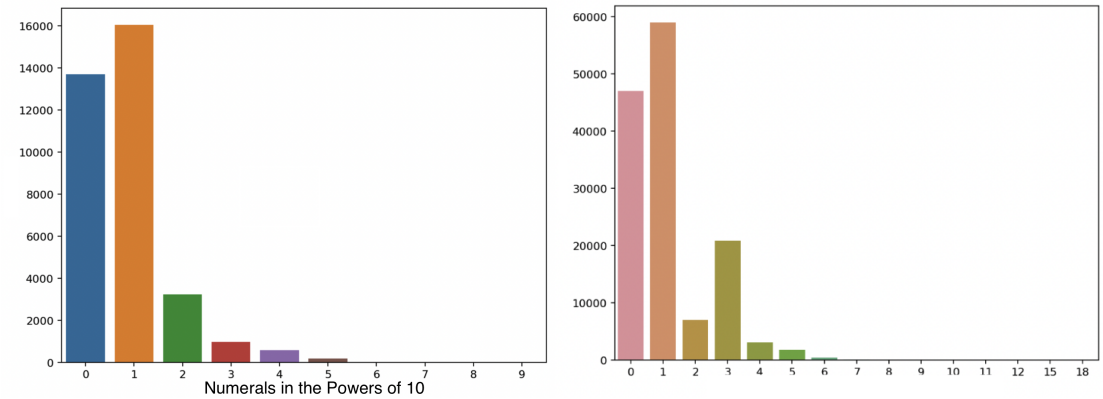
[27] Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. Do language embeddings capture scales? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 292–299, 2020.

5 Appendix

5.1 Limitations

We believe this work to be a proof-of-concept for addressing the important problem of non-linguistic skill injection in language modeling. Besides mathematical reasoning, the set of non-linguistic skills can expand to encompass logical inference, and dataset comprehension. Additionally, our in-house arithmetic dataset comprises of addition and subtraction - we hope to extend our highly generic proposed framework to incorporate other mathematical operations like multiplication, division, and exponentiation, in addition to incorporating decimal-point numerals in the dataset.

5.2 Numeral Distribution of the Arithmetic Dataset



(a) EQUATE (b) DROP

Figure 3: Numeral distribution histograms of the EQUATE & DROP datasets based on powers of 10.